Logistische Regression und Probit-Modelle mit SPSS: Anmerkungen zu zwei sehr unterschiedlichen Prozeduren

Hartmann, Peter H.

Logistische Regression und Probit-Modelle mit SPSS: Anmerkungen zu zwei sehr unterschiedlichen Prozeduren

Von Peter H. Hartmann

1. Problemstellung

Die Anwendung von Regressionsmodellen ist im Fall dichotomer abhängiger Variablen nicht ganz so einfach wie bei metrischen Merkmalen. [1] Wenn die Abwesenheit eines Zustands durch den Wert Null und das Vorhandensein dieses Zustands durch den Wert Eins gekennzeichnet wird, kann mit Hilfe eines Regressionsmodells die Wahrscheinlichkeit, daß dieser Zustand vorhanden ist, geschätzt werden. Wird diese Wahrscheinlichkeit mit Hilfe linearer Regression und des gewöhnlichen Kleinstquadratschätzers (OLS) geschätzt, dann ist die Schätzung der Regressionskoeffizienten nicht optimal, weil die Varianz einer dichotomen Variablen über die Beobachtungen hinweg nicht konstant ist.

Während sich dieses Problem noch relativ leicht, etwa durch ein gewichtetes Schätzverfahren (WLS), lösen läßt, ist ein zweites Problem der linearen Regression schwerwiegender: Für bestimmte Werte der unabhängigen Variablen kann es zu Schätzungen der abhängigen Variablen von weniger als Null oder mehr als Eins kommen. Interpretiert man die abhängige Variable als Wahrscheinlichkeit, dann ist dies nicht akzeptabel, denn Wahrscheinlichkeiten können per definitionem nicht kleiner als Null oder größer als Eins werden.

Einen Ausweg bietet hier die Wahl einer nichtlinearen Funktion, die in Abhängigkeit von den Werten der unabhängigen Variablen die Wahrscheinlichkeit angibt, daß der interessierende Zustand vorhanden ist. Diese Funktion wird so gewählt, daß die geschätzten Wahrscheinlichkeiten nie kleiner als Null oder größer als Eins werden können.

Das Vorhandensein des Zustands sei durch den Wert Eins einer dichotomen Variablen VDEP dargestellt, seine Abwesenheit durch den Wert Null. Die Wahrscheinlichkeit, daß der Zustand vorhanden ist, soll mit Hilfe der metrischen unabhängigen Variablen VINDEP1, VINDEP2, ..., VINDEPk

vorhergesagt werden. Unter diesen metrischen Variablen können auch 'Dummy'-codierte Merkmale mit den Ausprägungen Null und Eins sein.

Die Wahrscheinlichkeit, daß der Zustand vorliegt, wird nun üblicherweise dargestellt entweder durch die logistische Kurve

$$p \quad (VDEP = 1) \quad = \quad e^z/(1 + e^z)$$

oder alternativ durch die kumulierte Dichte der Normalverteilung:

$$p \quad (VDEP = 1) \quad = \quad \frac{1}{\sqrt{2\pi}} \quad \int_{-\infty}^{z} e^{-t^2/2} dt.$$

Im ersten Fall spricht man von logistischer Regression, im zweiten Fall vom Probit-Modell. Die Größe z steht dabei für die Summe der mit ihren Regressionsgewichten multiplizierten unabhängigen Variablen VINDEP1 ..., VINDEPk plus der Regressionskonstanten. Diese Regressionsgewichte und die Regressionskonstante sind dann zu schätzen.

Seit einiger Zeit enthält auch das weit verbreitete Paket SPSS Standardroutinen zur Schätzung von logistischen Regressionen und Probit-Modellen. In diesem Kontext sind insbesondere die Prozeduren LOGISTIC REGRESSION und PROBIT zu nennen.[2] Diese Programme erlauben Maximum-Likelihood-Schätzungen der interessierenden Regressionsgewichte. Jedoch sind die beiden Prozeduren für grundsätzlich andere Datenstrukturen angelegt, und so ergeben sich - besonders für den Anfänger - erhebliche Schwierigkeiten beim Vergleich der geschätzten Koeffizienten.

2. Logistische Modelle mit der SPSS-Prozedur LOGISTIC REGRESSION

Bei SPSS und SPSS-PC (ab Version 3.1), nicht jedoch bei SPSS-X, können mit dem Programm LOGISTIC REGRESSION logistische Regressionsmodelle mit dichotomen abhängigen Variablen geschätzt werden.[3] Die Schätzung von Probit-Modellen ist mit dieser Prozedur dagegen nicht möglich.

2.1 Individualdaten

Betrachten wir zunächst Daten auf Individualebene. Abbildung 1 zeigt beispielhaft eine Datei auf der Ebene von Individualdaten.

Abbildung 1: Individualdaten

VDEP	KONST	(Diverse unabh. Variablen)
0	1	...
1	1	...
1	1	...
0	1	...
1	1	...
0	1	...
.	.	.
.	.	.
.	.	.

Abbildung 2: Aggregierte Daten: Kombinationen aller Variablen

	(Diverse Kombination der abhängigen und aller unabhängigen Variablen)	
FZ	VDEP	
5	0	1.Kombination der unabh. Variablen
3	1	1.Kombination der unabh. Variablen
4	0	2.Kombination der unabh. Variablen
2	1	2.Kombination der unabh. Variablen
3	0	3.Kombination der unabh. Variablen
3	1	3.Kombination der unabh. Variablen
.	.	.
.	.	.
.	.	.

In diesem Fall können die Koeffizienten des logistischen Regressionsmodells durch die folgende Anweisung geschätzt werden.

LOGISTIC REGRESSION VDEP WITH VINDEP1, VINDEP2, ..., VINDEPk

Bei der Liste der unabhängigen Variablen ist zu beachten, daß das Kommando LOGISTIC REGRESSION die Abkürzung mit 'TO' nicht akzeptiert. Anstelle von VINDEP1, VINDEP2, ..., VINDEPk kann man also nicht schreiben VINDEP1 TO VINDEPk. Alle Variablennamen müssen einzeln eingegeben werden.

Bei größerer Zahl der Fälle oder Tabellenfelder empfiehlt sich bei SPSS-PC eine Auslagerung der Berechnungen aus dem Kernspeicher, die mit der Option

/EXTERNAL

zu erreichen ist. Weiterhin müssen bei der Anwendung von LOGISTIC REGRESSION auf große Datenmengen ggf. lange Rechenzeiten in Kauf genommen werden.

2.2 Aggregierte Daten: Kombinationen aller Variablen

Bei sehr großen Fallzahlen ist es zur Einsparung von Speicherbedarf und Rechenzeit sinnvoll, die Daten zu aggregieren. Dies ist vor allem angebracht, wenn die metrischen unabhängigen Variablen nur wenige Ausprägungen haben (Wackerow 1990). Üblicherweise werden Variablen so voraggregriert, daß jeder besetzten Kombinationsmöglichkeit aller (d.h. aller abhängigen und aller unabhängigen) Variablen genau ein Satz entspricht. Dieser Satz enthält neben den jeweiligen Ausprägungen der abhängigen und aller unabhängigen Variablen eine weitere Variable, die die Häufigkeit der Merkmalskombination angibt, den sogenannten Fallzähler. Wir bezeichnen ihn mit FZ.

Die Voraggregation kann z.B. unter der Verwendung der Prozedur AGGREGATE geschehen:[4)]

```
AGGREGATE OUTFILE=*
 /BREAK = VDEP VINDEP1, ..., VINDEPk
 /FZ   = N 0
```

Da AGGREGATE nur speicherresident sortieren kann, sollte bei großen Fall- und Variablenzahlen die Daten mit SORT CASES vorsortiert werden, bei

AGGREGATE ist dann die OPTION '/PRESORTED' anzugeben. SORT CASES sortiert bei nicht hinreichender Kernspeicherallokation unter Verwendung von Hilfsdateien, die ggf. angeschlossen werden müssen (z.B. bei SIEMENS-BS 2000 unter dem Linknamen IOSDA).

Durch die Aggregation entsteht eine Datei, deren Aufbau als Beispiel in Abbildung 2 dargestellt wird.

Bei so aggregierten Daten muß vor Anwendung der Schätzprozedur gewichtet werden, und zwar wie generell üblich mit dem Kommando WEIGHT:

WEIGHT BY FZ.

3. Logistische und Probit-Modelle mit der SPSS-Prozedur PROBIT

Will man mit Hilfe von SPSS nicht nur logistische, sondern auch PROBIT-Modelle schätzen, so ist man auf die Prozedur PROBIT angewiesen. Diese steht sowohl bei SPSS als auch bei SPSS-PC (ab Version 4.0) zur Verfügung, ebenfalls bei SPSS-X (ab Version 2.1). Dabei sind allerdings einige Besonderheiten bei den Datenformaten zu berücksichtigen.

3.1 Individualdaten

Üblicherweise erwartet die Prozedur PROBIT die Daten in einer speziellen, aggregierten Form. Auf diese wird in Abschnitt 3.2 eingegangen. Liegen die Daten auf Einzelfallebene vor, ist zunächst eine Konstante mit dem Wert Eins zu berechnen:

COMPUTE KONST=1

Als Prozedurkommando ist dann zu schreiben:

PROBIT VDEP OF KONST WITH VINDEP1, VINDEP2, ..., VINDEPk

SPSS PROBIT berechnet in diesem Fall korrekte Regressionskoeffizienten und Standardfehler. Falsch dagegen ist der Chi-Quadrat-Anpassungstest.

In der Praxis stößt die Anwendung von SPSS PROBIT auf Einzelfalldaten bei großer Fall- und Variablenzahl schnell an Grenzen, weil der Hauptspeicherbedarf der Prozedur nicht nur mit der Variablenzahl, sondern auch mit der

Fallzahl zunimmt und, abweichend von der Prozdur LOGISTIC REGRESSION, keine Möglichkeit der externen Zwischenspeicherung besteht.

3.2 Aggregierte Daten: Kombinationen aller unabhängigen Variablen

Von seiner Konzeption her ist das Programm PROBIT für die Schätzung von Dosis-Response Modellen optimiert. Norusis/SPSS Inc. (1990a:227) geben als Beispiele Fragestellungen an wie: "Bei welcher Dosis eines Insektizids stirbt welcher Anteil an Insekten?" oder: "Bei welchem Preis kauft welcher Anteil potentieller Käufer ein bestimmtes Produkt?". Die zu schätzenden Anteile sind aus dichotomen Individualmerkmalen aggregierte Größen: Ein Insekt stirbt oder stirbt nicht; ein Käufer kauft oder kauft nicht.

Liegen Individualdaten vor, sollte man sie für die Eingabe bei PROBIT also nach Möglichkeit aggregieren. PROBIT erwartet dabei aber ein anderes Datenformat als LOGISTIC REGRESSION. Während bei LOGISTIC REGRESSION die Aggregation so erfolgen sollte, daß jeder besetzten Kombinationsmöglichkeit aller (d.h. der abhängigen und aller unabhängigen) Variablen genau ein Satz entspricht (vgl. Abbildung 2), erwartet PROBIT die Daten so, daß jeder besetzten Kombinationsmöglichkeit aller unabhängigen (nicht jedoch der abhängigen) Variablen genau ein Satz entspricht. Dieser Satz enthält neben den jeweiligen Ausprägungen aller unabhängigen Variablen zwei weitere Variable. Die eine (FZ) ist wiederum ein Fallzähler, der die Häufigkeit der jeweiligen Kombination anzeigt. Die andere (FVDEP) enthält nun die Häufigkeit der Fälle, bei denen - gegeben die jeweilige Kombination der unabhängigen Variablen - die abhängige Variable den Wert Eins annimmt.

Bei den oben genannten Beispielen steht FZ also für die Zahl der Insekten oder die Zahl potentieller Käufer, FVDEP dagegen für die Zahl der gestorbenen Insekten oder die Zahl der tatsächlichen Käufer.
Eine solche Voraggregation kann wiederum z.B. unter der Verwendung der Prozedur AGGREGATE geschehen. Liegen vor der Aggregation Individualdaten (Abbildung 1) vor, so ist zu schreiben:

```
AGGREGATE OUTFILE=*
 /BREAK = VINDEP1, ..., VINDEPk
 /FZ    = N  (VDEP)
 /FVDEP = SUM (VDEP)
```

Liegen dagegegen bereits aggregierte Daten in der in Abbildung 2 geschilderten Form vor, so ist zu schreiben:

```
COMPUTE FVDEP=FZ*VDEP
AGGREGATE OUTFILE=*
 /BREAK = VINDEP1, ..., VINDEPk
 /FZ    = SUM (FZ)
 /FVDEP = SUM (FVDEP)
```

Abbildung 3: Aggregierte Daten: Kombinationen aller unabhängigen Variablen

FVDEP	FZ	(Diverse Kombinationen aller unabhängigen Variablen)
3	8	1.Kombination der unabh. Variablen
2	6	2.Kombination der unabh. Variablen
3	6	3.Kombination der unabh. Variablen
.	.	.
.	.	.
.	.	.

Abbildung 4: Aggregierte Daten: Kombinationen aller Variablen und Variable FVDEP

FVDEP	FZ	(Diverse Kombination der abhängigen und aller unabhängigen Variablen) VDEP	
0	5	0	1.Kombination der unabh. Variablen
3	3	1	1.Kombination der unabh. Variablen
0	4	0	2.Kombination der unabh. Variablen
2	2	1	2.Kombination der unabh. Variablen
0	3	0	3.Kombination der unabh. Variablen
3	3	1	3.Kombination der unabh. Variablen
.	.	.	.
.	.	.	.
.	.	.	.

Der Aufbau der zu analysierenden aggregierten Datei wird in Abbildung 3 dargestellt.

Als Prozedurkommando für die Analyse ist einzugeben:

PROBIT FVDEP OF FZ WITH VINDEP1, VINDEP2, ..., VINDEPk

Wie im Fall der Eingabe von Individualdaten berechnet SPSS PROBIT korrekte Regressionskoeffizienten und Standardfehler. Der Chi-Quadrat-Anpassungstest ist bei dieser Form der Dateneigabe - und nur bei dieser Form (!) - korrekt. Eine Gewichtung mit dem Kommando WEIGHT ist nicht erforderlich und wird vom Kommando PROBIT auch ignoriert.

3.3 Aggregierte Daten: Kombinationen aller Variablen

Aggregierte Daten, bei denen jeder besetzten Kombinationsmöglichkeit aller (d.h. der abhängigen und aller unabhängigen) Variablen genau ein Satz entspricht (vgl. Abbildung 2), können alternativ zum eben geschilderten Vorgehen auch ohne weitere Aggregation direkt mit PROBIT analysiert werden. Hierbei ist ähnlich wie in Abschnitt 3.1 (Individualdaten) vorzugehen. Anstelle der Konstanten wird nun der Fallzähler FZ eingetragen. Zusätzlich ist aber anstelle der abhängigen Variable deren Tabellenhäufigkeit einzutragen. Diese Häufigkeit ergibt sich wie in Abschnitt 3.2 aus der Multiplikation der zu erklärenden Dichotomie mit dem Fallzähler.

*COMPUTE FVDEP = FZ * VDEP*

Die hier zu analysierende Datei hätte dann etwa einen Aufbau wie Abbildung 4.

Als Prozedurkommando ist zu geben:

PROBIT FVDEP OF FZ WITH VINDEP1, VINDEP2, ..., VINDEPk

SPSS PROBIT berechnet in diesem Fall bei Dateneingabe auf Individualebene korrekte Regressionskoeffizienten und Standardfehler, aber einen falschen Chi-Quadrat-Anpassungstest.

Die Verwendung von FVDEP als abhängige Variable anstelle von VDEP ist wichtig! Verwendet man nämlich bei der gegebenen Datenstruktur VDEP als abhängige Variable, so kommt man zu falschen, möglicherweise aber halbwegs plausiblen Regressionskoeffizienten. Falsch ist also insbesondere bei der hier gegebenen Datenstruktur die Spezifikation:

PROBIT VDEP OF FZ WITH VINDEP1, VINDEP2, ..., VINDEPk

3.4 Optionen

Einige bei SPSS PROBIT voreingestellte Optionen sind für logistische oder Probit-Regressionen nicht sinnvoll.

So sollte für logistische oder Probit-Regressionen im Regelfall die als Standard voreingestellte Logarithmierung aller unabhängigen Variablen mit der Option

/LOG=NONE

abgestellt werden.

Weiterhin ist es bei großen Fall- oder Zellenzahlen sinnvoll, die als Standard voreingestellte Ausgabe der vorhergesagten Werte, Schätzfehler etc. für die Einzelfälle (Datenstrukur 3.1) oder Tabellenfelder (Datenstruktur 3.2 oder 3.3) zu unterdrücken, da sonst sehr umfangreiche Drucklisten erzeugt werden. Durch die Angabe von

/PRINT=CI RMP

werden die voreingestellten Druckoptionen mit Ausnahme der umfangreichen Einzelfallstatistiken ausgegeben.

3.5 Transformation der Koeffizienten

Bei der Interpretation der Koeffizienten von PROBIT ist zu berücksichtigen, daß SPSS die Koeffizienten in etwas unüblicher Art und Weise transformiert.

So wird zu dem Koeffizienten für die Regressionskonstante (Intercept) von SPSS der Wert von Fünf addiert. Um den korrekten Koeffizienten zu erhalten, ist Fünf vom ausgedruckten Wert abzuziehen.

Weiterhin dividiert SPSS bei der Option '/MODEL=LOGIT' die Regressionskoeffizienten durch Zwei. Dies geschieht, damit die vom Logit-Modell ausgegebenen Koeffizienten eine ähnliche Größenordnung haben wie die des Probit-Modells. Um jedoch die korrekten Koeffizienten der logistischen Regression zu erhalten (die identisch mit den Koeffizienten sind, die vom Programm LOGISTIC REGRESSION ausgegeben werden), ist der ausgedruckte Wert mit Zwei zu multiplizieren.

Die korrekten Regressionskoeffizienten b sind also je nach Modell aus den von der Prozedur PROBIT ausgedruckten Koeffizienten wie folgt zu berechnen:

a) Option: */MODEL=LOGIT*:
   Konstante: b = (b - 5) * 2
   übrige Koeffizienten: b = b * 2

b) Option: */MODEL=PROBIT*:
   Konstante: b = b - 5
   übrige Koeffizienten: b = b

4. Fazit

Mit den Prozeduren LOGISTIC REGRESSION und PROBIT bietet SPSS leistungsfähige Produkte zur logistischen und zur Probit-Analyse von binären abhängigen Variablen an.

Wenn nur logistische Regressionsgewichte und keine Probit-Schätzungen benötigt werden, sollten im Regelfall die Analysen mit dem Programm LOGISTIC REGRESSION durchgeführt werden, da dies durch die übersichtlicheren Analyse- und vielfältigeren Ausgabemöglichkeiten gekennzeichnet ist.

Sind dagegen auch Probit-Schätzungen erforderlich, so muß der Nutzer von SPSS auf das Programm PROBIT umsteigen. Da dieses Programm eine andere Datenstruktur als Eingabe erwartet und da es keine externe Zwischenspeicherung von Matrizen erlaubt, sollten umfangreichere Daten vor der Analyse grundsätzlich aggregiert werden. Bei der Aggregation sind die besonderen, in den Abschnitten 3.2 und 3.3 dieses Artikels beschriebenen Eingabevoraussetzungen für PROBIT zu beachten. Die Logarithmierung aller unabhängigen Variablen ist abzuschalten, ebenfalls sollte nicht grundsätzlich die voreingestellte Einzelfalldiagnostik ausgegeben werden. Schließlich sollten die Ergebnisse in der in Abschnitt 3.5 geschilderten Weise wieder in die ursprünglichen Ergebnisse zurücktransformiert werden.

Insgesamt erscheint es wenig verständlich, weshalb SPSS zwei hinsichtlich ihrer Eingabedaten derart unterschiedliche Programme für so ähnliche Probleme anbietet. Für die meisten sozialwissenschaftlichen Nutzer wäre es hilfreicher, wenn das Programm LOGISTIC REGRESSION als eine Option auch PROBIT-Regressionen schätzen könnte.

Anmerkungen

1) Eine leicht verständliche Einführung in die Probleme der Regression bei dichotomen oder polytomen abhängigen Variablen geben Aldrich/Nelson (1984). Wichtige Hinweise zu diesem Aufsatz gab Siegfried Gabler.
2) Die hier diskutierten Prozeduren sind dokumentiert in Norusis/SPSS Inc. (1990). Praktische Beispiele zu ihrer Anwendung finden sich in Norusis/SPSS Inc. (1990a).
3) Mit Hilfe der SPSS-Standardprozeduren können nur binäre logistische Regressionen geschätzt werden. Das heißt die abhängige Variable ist eine Dichotomie und keine polytome Variable. Wie man mit SPSS (Prozedur MATRIX) auch logistische Regressionen bei polytomen abhängigen Variablen schätzen kann, zeigt Kühnel (1990). Probit-Modelle sind für polytome abhängige Variablen aus rechentechnischen Gründen in der Regel nicht schätzbar.
4) Nicht eingegangen wird hier auf das Problem fehlender Werte. Bei den Aggregationen ist sicherzustellen, daß Fälle mit fehlenden Werten entweder bereits vor oder bei der Aggregation ausgeschlossen werden, oder daß die fehlenden Werte nach der Aggregation wieder als MISSING markiert werden.

Literatur

Aldrich, J.H./Nelson, F.D., 1984: Linear Probability, Logit, and Probit Models. Beverly Hills: Sage.

Kühnel, S., 1990: Lassen sich mit SPSSx-Matrix anwenderspezifische Analyseprobleme lösen? Ein Anwendungstest am Beispiel der multinomialen logistischen Regression. ZA-Information 27:89-109.

Norusis, M.J./SPSS Inc., 1990: SPSS Reference Guide. Chicago.

Norusis, M.J./SPSS Inc., 1990a: SPSS Advanced Statistics User's Guide. Chicago.

Wackerow, J., 1990: Verarbeitung von Dateien mit großen Fallzahlen. S.534-541 in: F.Faulbaum/R.Haux/K.-H.Jöckel (Hrsg.), Fortschritte der Statistik-Software 2. SOFTSTAT '89. 5.Konferenz über die wissenschaftliche Anwendung von Statistik-Software, Heidelberg 1989. Stuttgart, New York: Fischer.